Dimensionality reduction distills complex evolutionary relationships in seasonal influenza and SARS-CoV-2

Sravani Nanduri[1], Allison Black[2], Trevor Bedford[2,3], John Huddleston[2*]

1 Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA

2 Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA

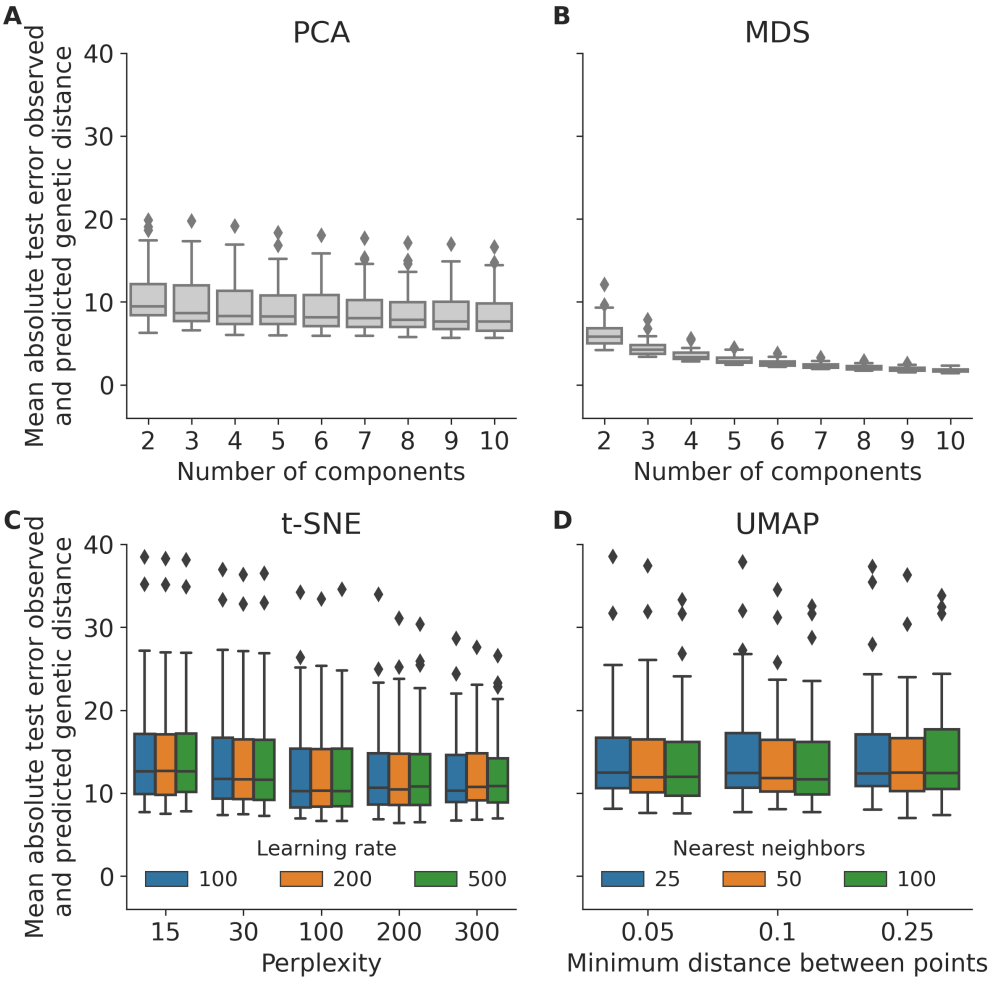
3 Howard Hughes Medical Institute, Seattle, WA, USA

* jhuddles@fredhutch.org

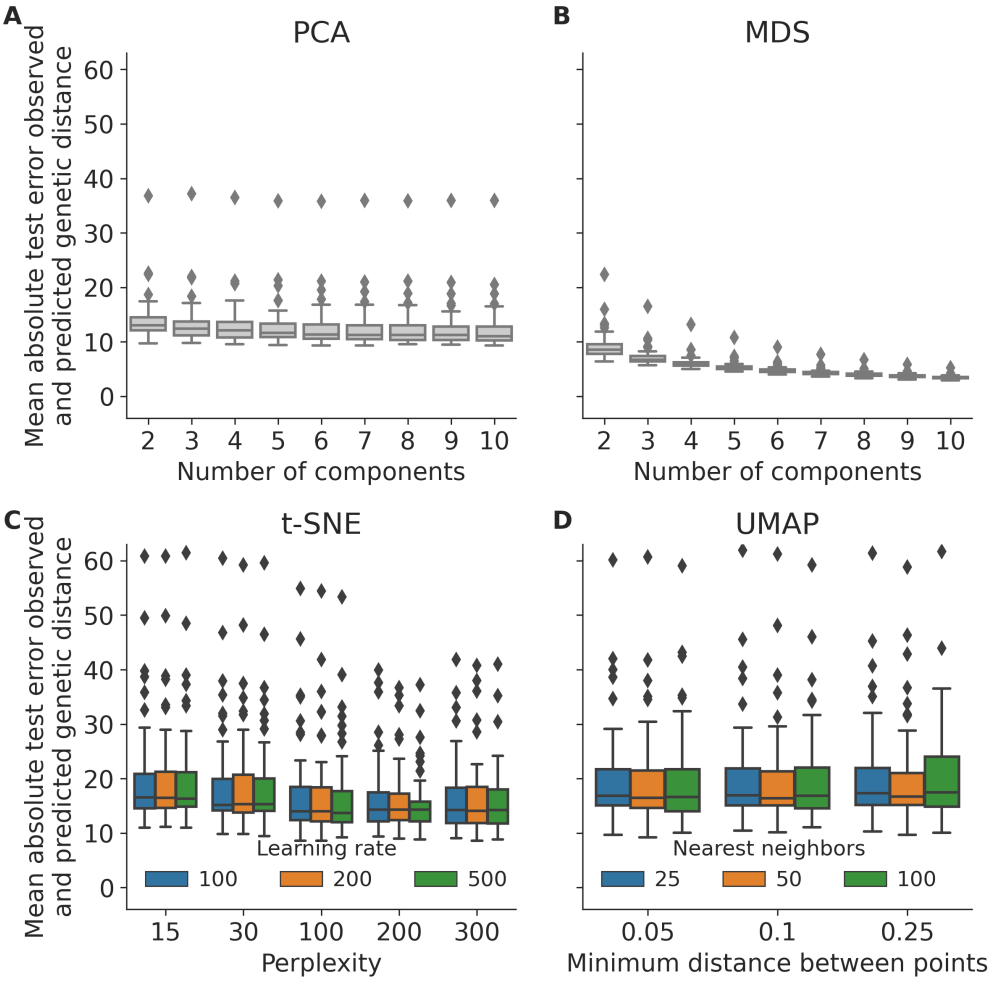
Supporting Information

S1 Table. Optimal cluster thresholds per pathogen, known genetic group type, and embedding method based on normalized variation of information (VI) distances calculated from early pathogen datasets. Smaller VI values indicate fewer differences between HDBSCAN clusters and known genetic groups. VI of 0 indicates identical clusters and 1 indicates maximally different clusters. Threshold refers to the minimum Euclidean distance between initial clusters for HDBSCAN to consider them as distinct clusters. We apply these optimal thresholds per pathogen, known genetic group type, and method to find clusters in corresponding late datasets for each pathogen.

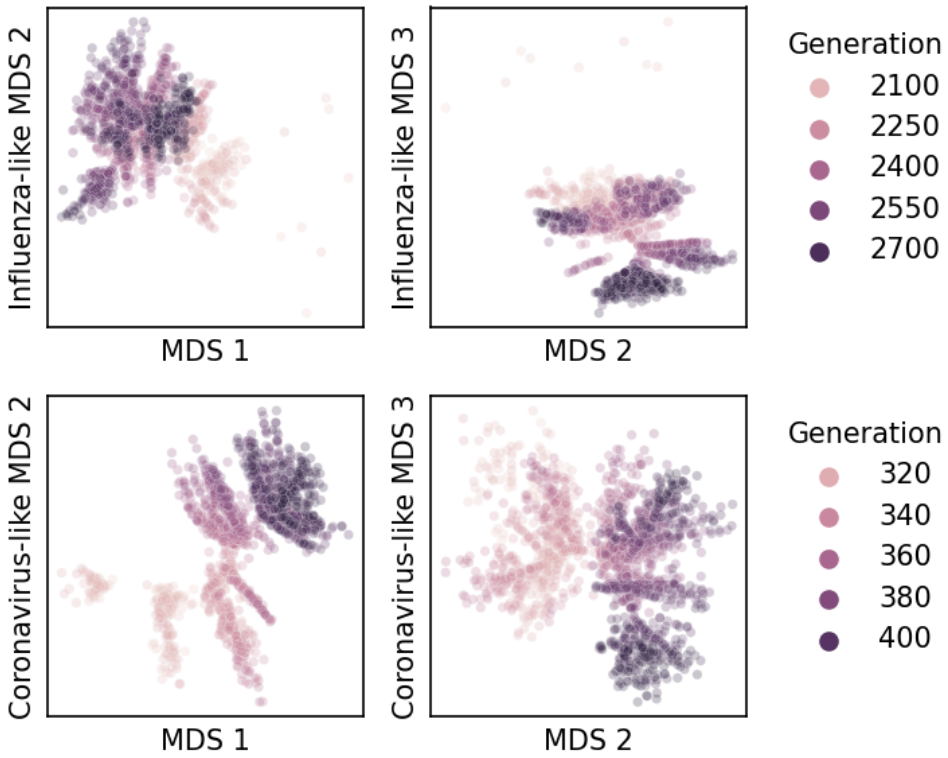
Pathogen	Genetic Group Type	Method	VI	Threshold
Influenza H3N2	Nextstrain clade	t-SNE	0.04	2.0
		UMAP	0.09	1.0
		MDS	0.11	3.5
		PCA	0.19	0.5
SARS-CoV-2	Nextstrain clade	t-SNE	0.07	1.0
		MDS	0.15	0.0
		UMAP	0.16	0.5
		PCA	0.22	0.5
SARS-CoV-2	Pango	t-SNE	0.12	1.0
		MDS	0.23	0.0
		UMAP	0.25	0.5
		PCA	0.31	0.5



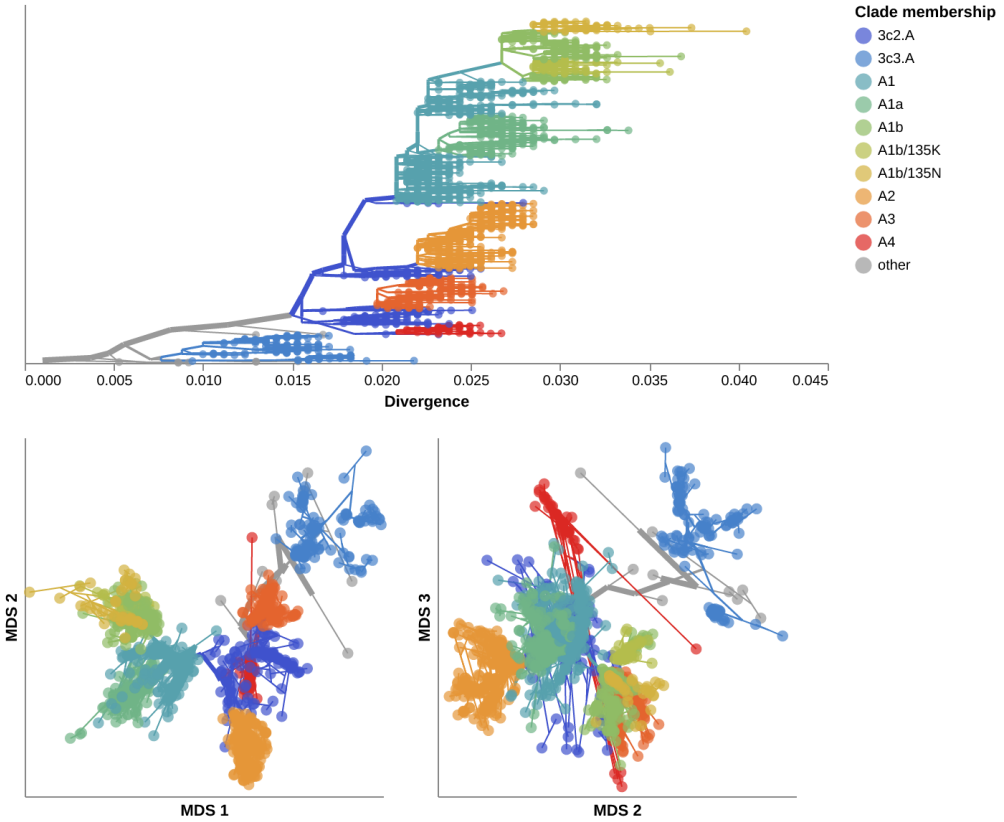
S1 Fig. Distribution of mean absolute errors (MAE) between observed and predicted pairwise genetic distances per embedding method parameters for simulated influenza-like populations.



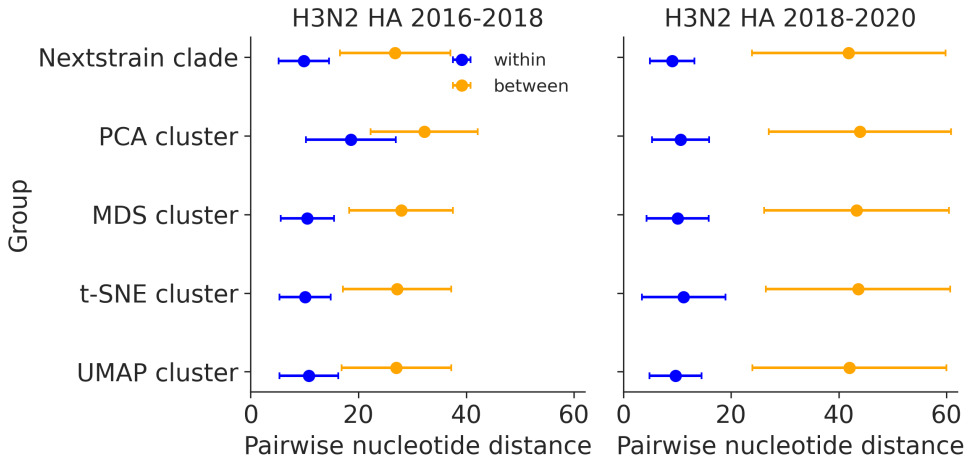
S2 Fig. Distribution of mean absolute errors (MAE) between observed and predicted pairwise genetic distances per embedding method parameters for simulated coronavirus-like populations.



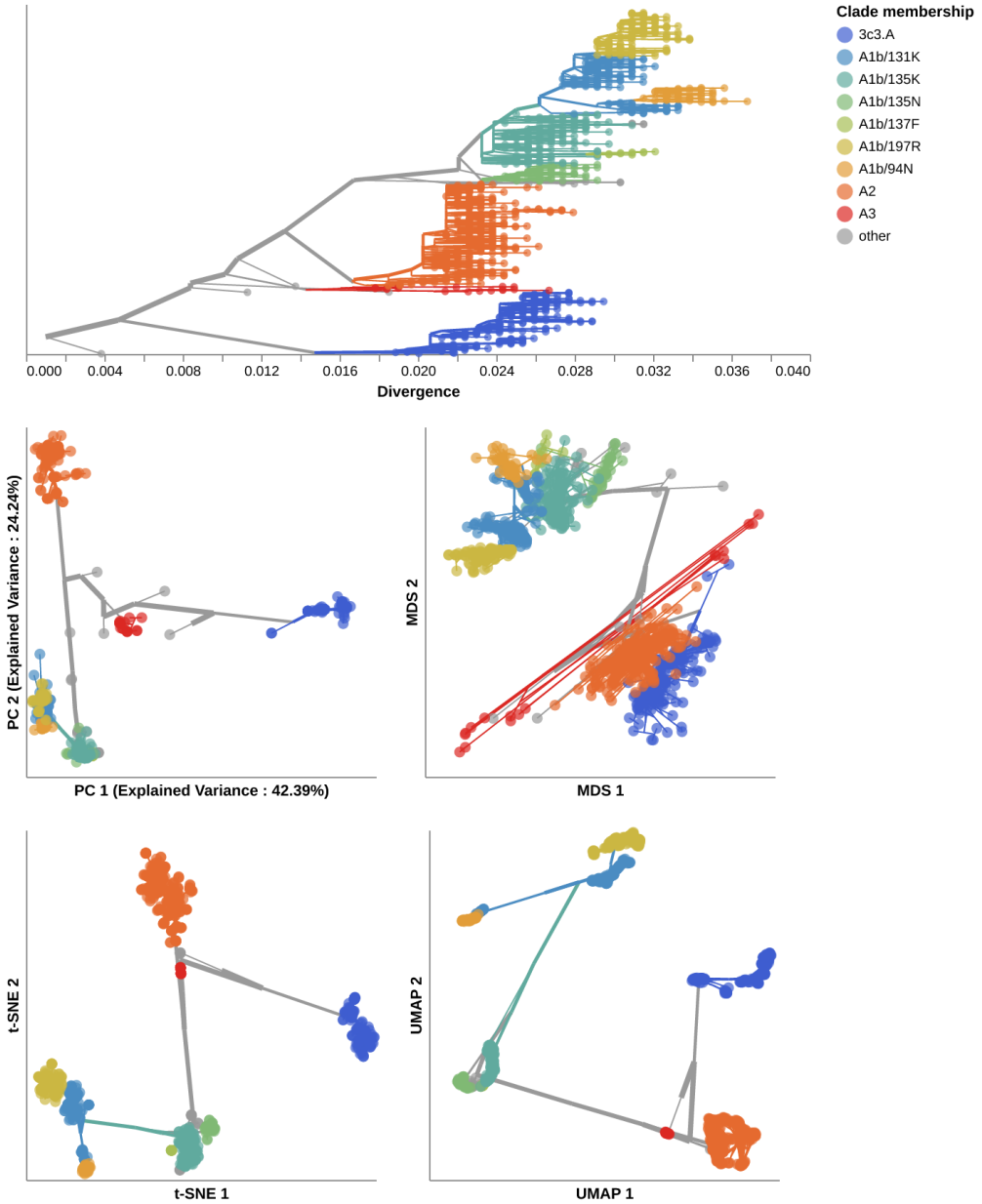
S3 Fig. Representative MDS embeddings for simulated populations using optimal parameters per pathogen (rows) and showing all three components.



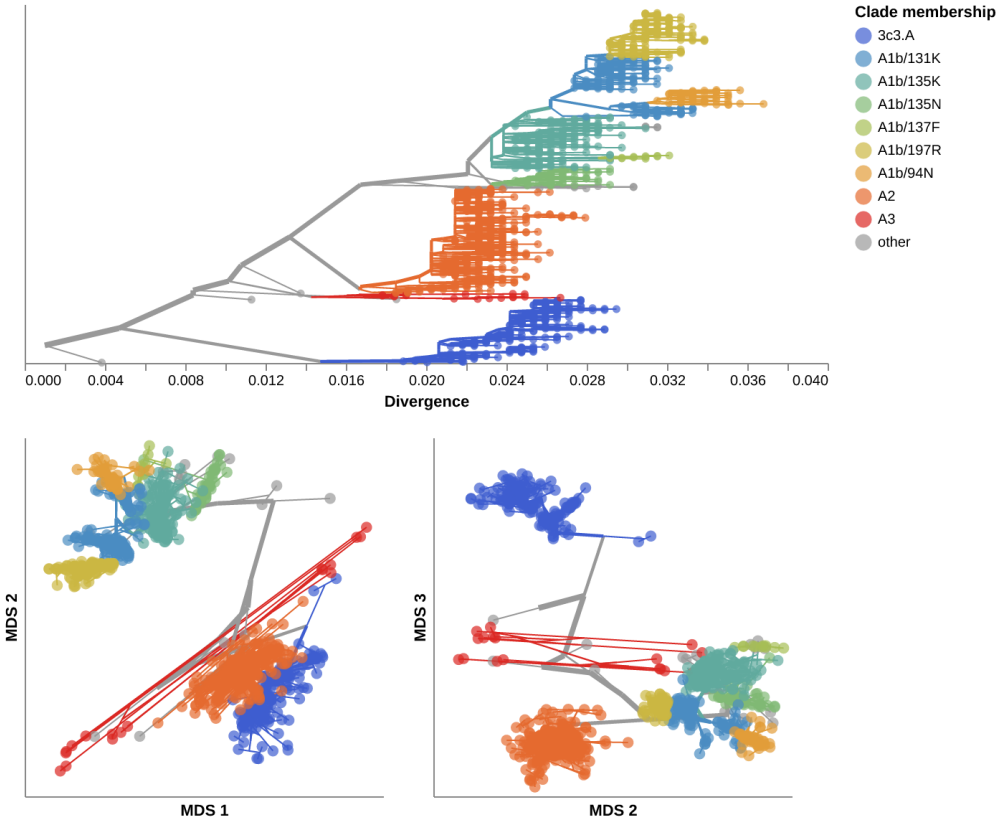
S4 Fig. MDS embeddings for early (2016–2018) influenza H3N2 HA sequences showing all three components. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line colors represent the clade membership of the most ancestral node in the pair of nodes connected by the segment. Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.



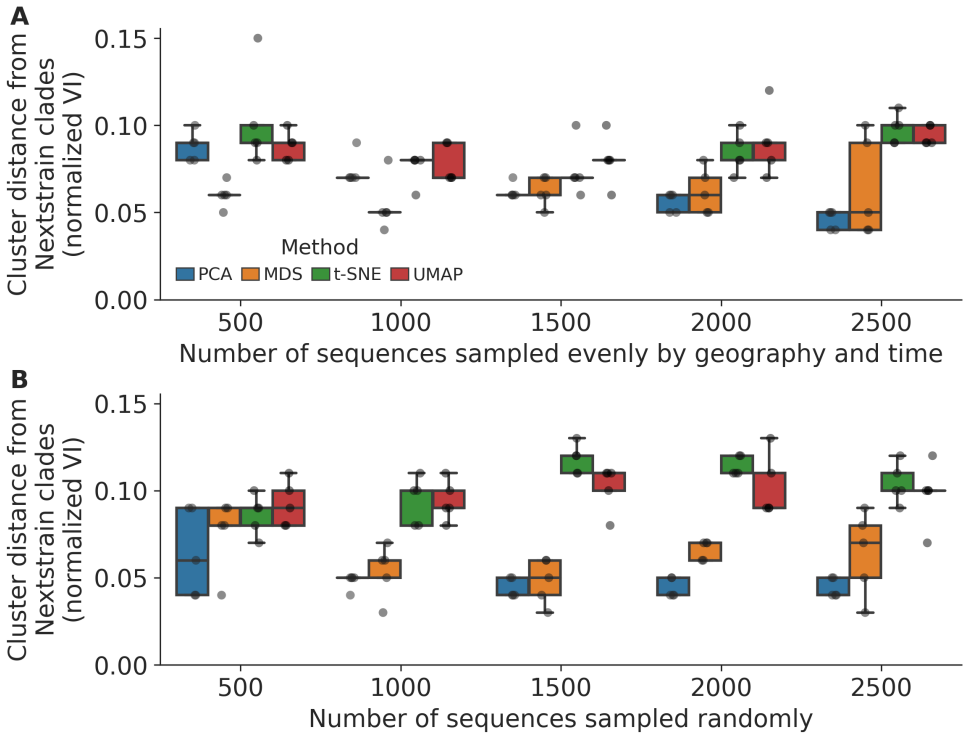
S5 Fig. Pairwise nucleotide distances for early (2016–2018) and late (2018–2020) influenza H3N2 HA sequences within and between genetic groups defined by Nextstrain clades and clusters from PCA, MDS, t-SNE, and UMAP embeddings.



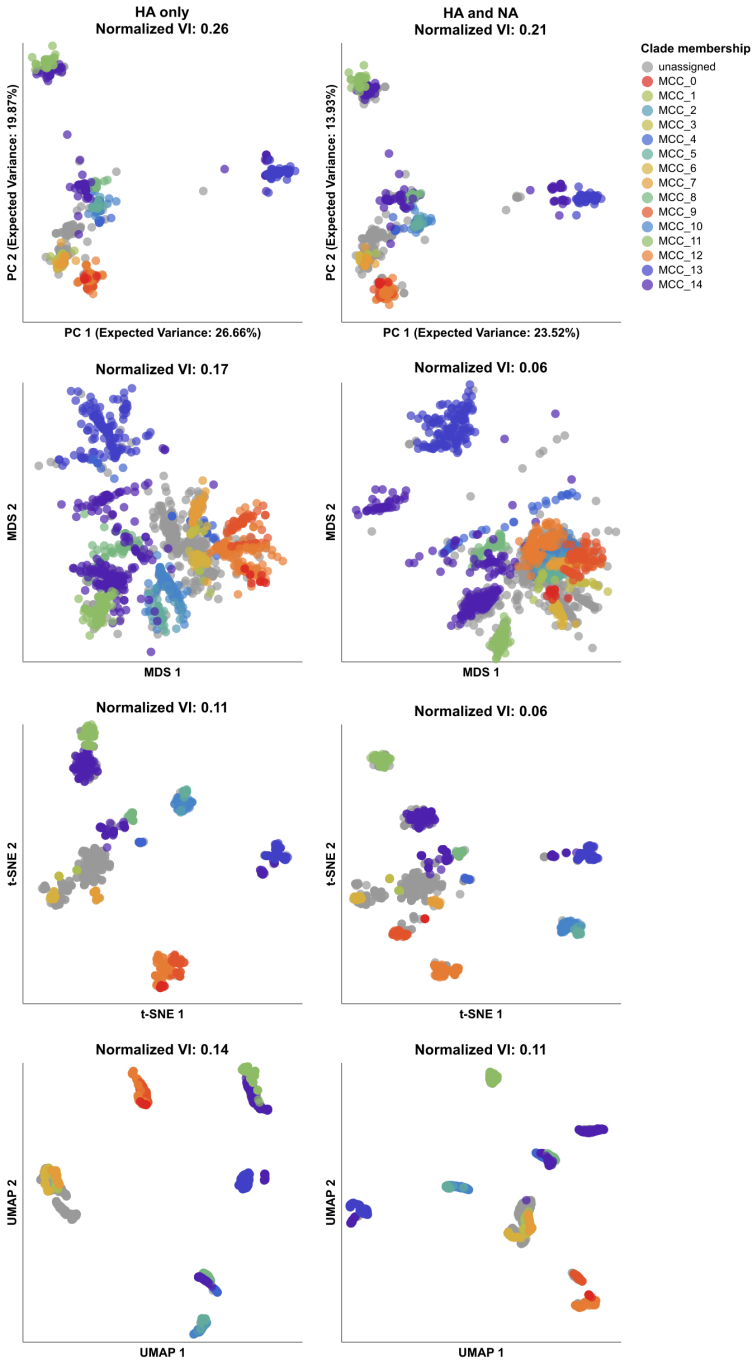
S6 Fig. Phylogeny of late (2018–2020) influenza H3N2 HA sequences plotted by nucleotide substitutions per site on the x-axis (top) and low-dimensional embeddings of the same sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their Nextstrain clade assignment. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line colors represent the clade membership of the most ancestral node in the pair of nodes connected by the segment. Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.



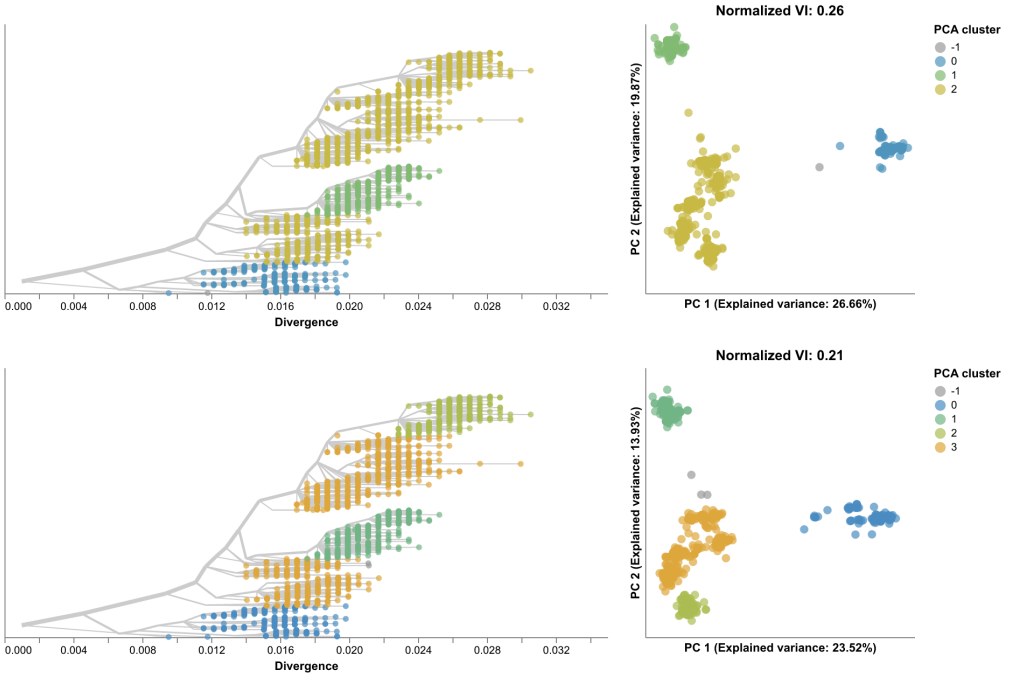
S7 Fig. MDS embeddings for late (2018–2020) influenza H3N2 HA sequences showing all three components. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line colors represent the clade membership of the most ancestral node in the pair of nodes connected by the segment. Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.



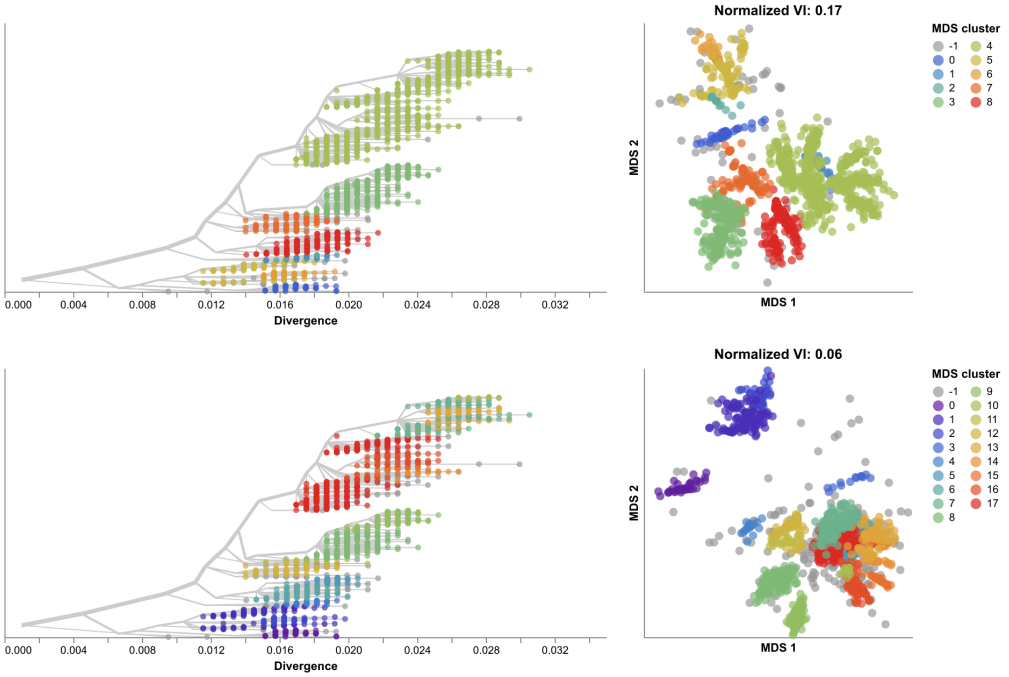
S8 Fig. Replication of cluster accuracy per embedding method for late (2018–2020) influenza H3N2 HA sequences across different sampling densities (total sequences sampled) and sampling schemes including A) even geographic and temporal sampling and B) random sampling. We measured cluster accuracy across five replicates per sampling density and scheme with the normalized VI distance between clusters from a given embedding and Nextstrain clades for the same samples. The even sampling scheme selected sequences evenly across country, year, and month to minimize geographic and temporal bias. The random sampling scheme uniformly sampled from the original dataset, reflecting the geographic and genetic bias in those data.



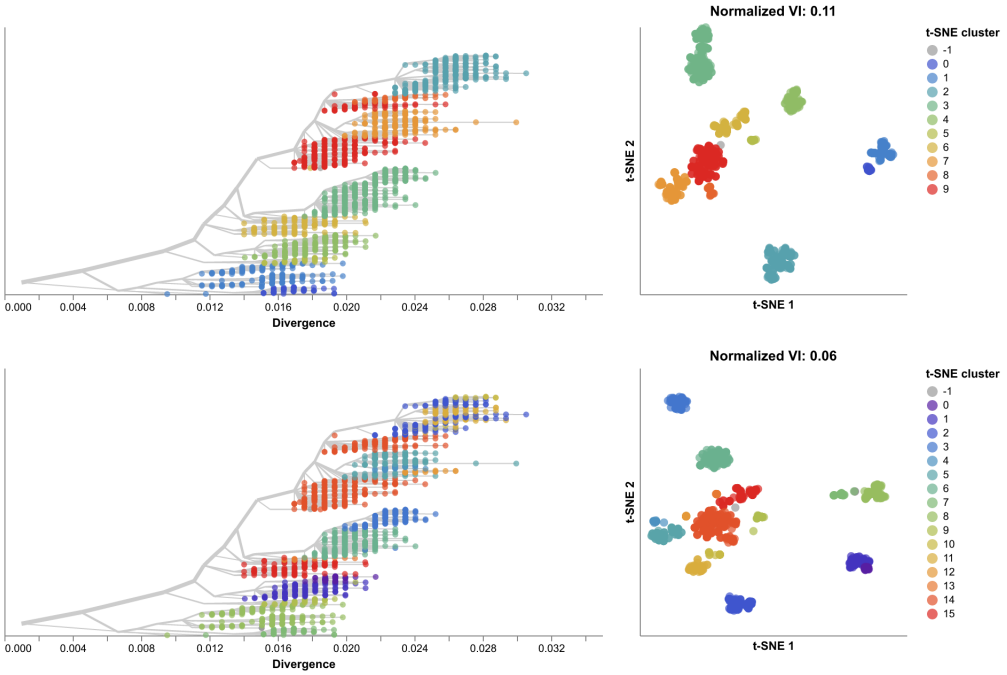
S9 Fig. Embeddings influenza H3N2 HA-only (left) and combined HA/NA (right) showing the effects of additional NA genetic information on the placement of reassortment events detected by TreeKnit (MCCs).



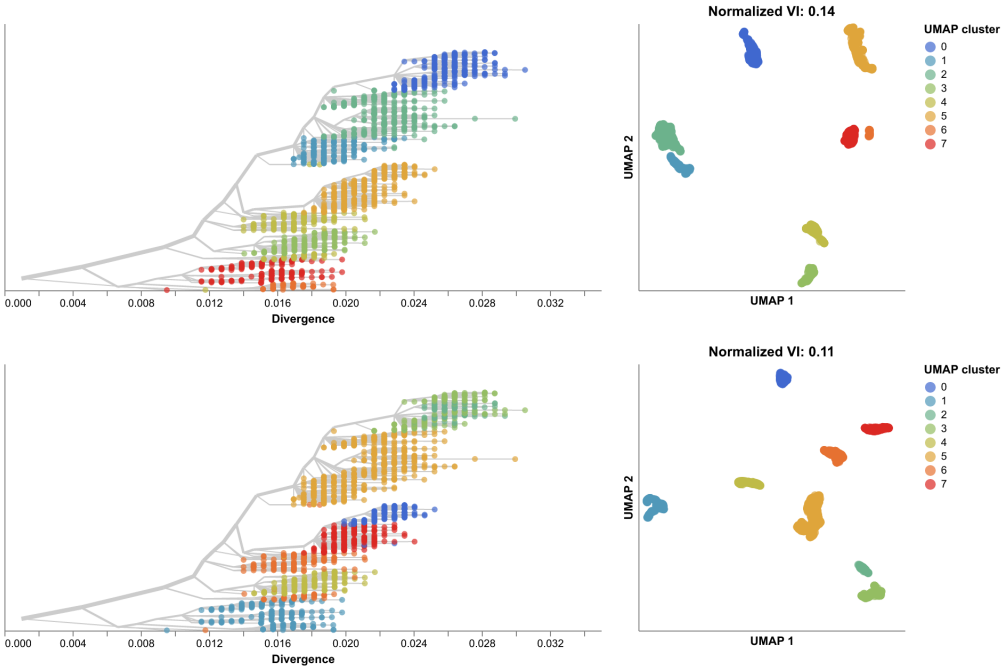
S10 Fig. PCA embeddings for influenza H3N2 HA sequences only (top row) and HA/NA sequences combined (bottom row) showing the HA trees colored by clusters identified in each embedding (left) and the corresponding embeddings colored by cluster (right).



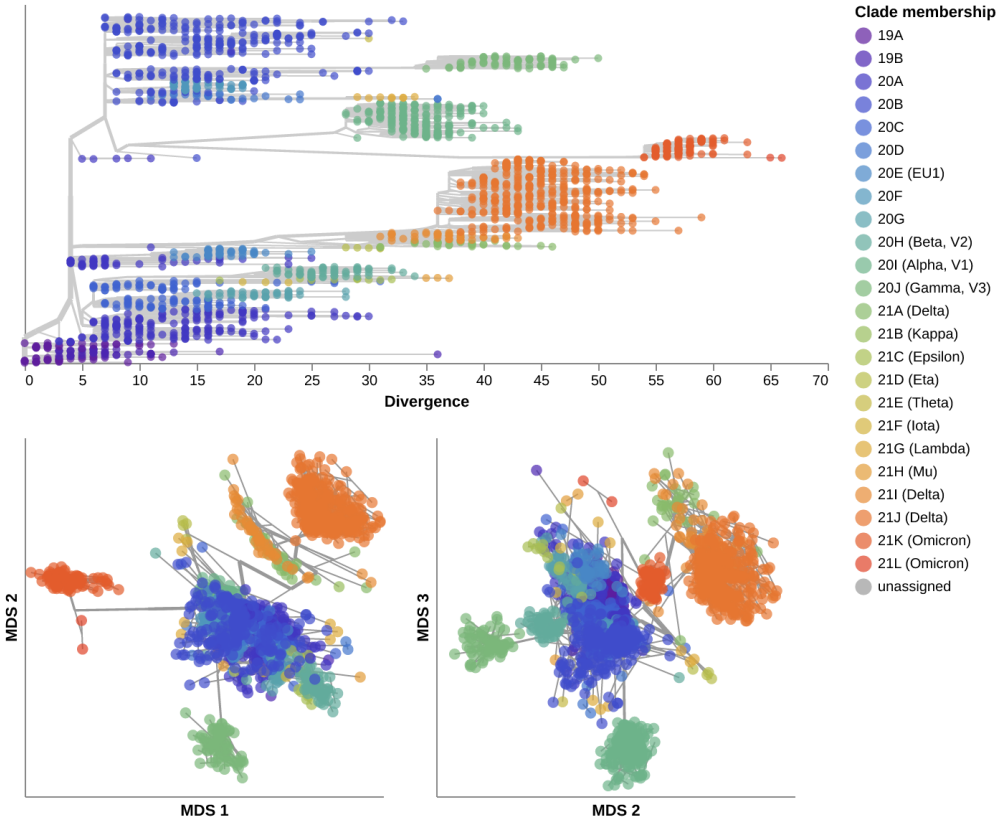
S11 Fig. MDS embeddings for influenza H3N2 HA sequences only (top row) and HA/NA sequences combined (bottom row) showing the HA trees colored by clusters identified in each embedding (left) and the corresponding embeddings colored by cluster (right).



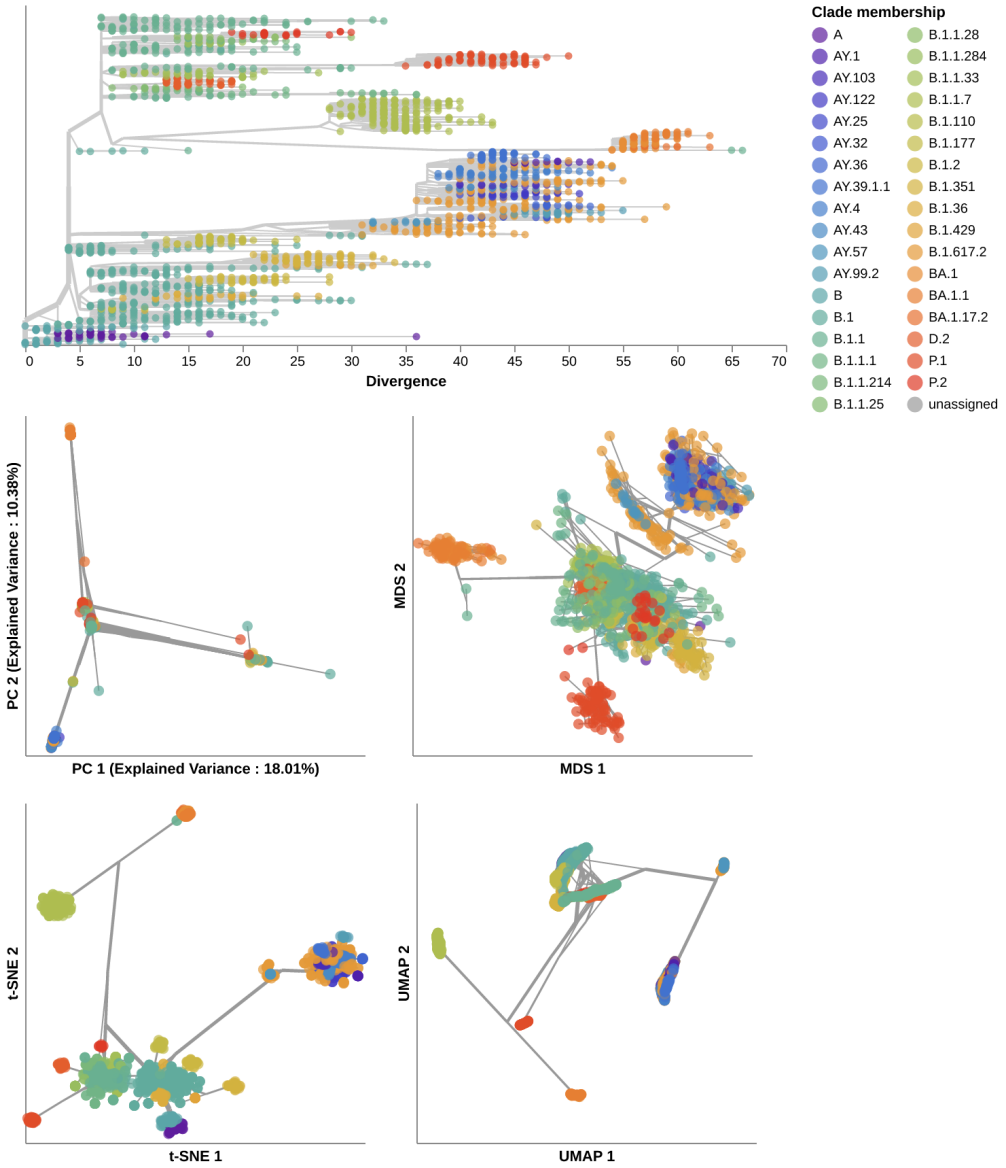
S12 Fig. t-SNE embeddings for influenza H3N2 HA sequences only (top row) and HA/NA sequences combined (bottom row) showing the HA trees colored by clusters identified in each embedding (left) and the corresponding embeddings colored by cluster (right).



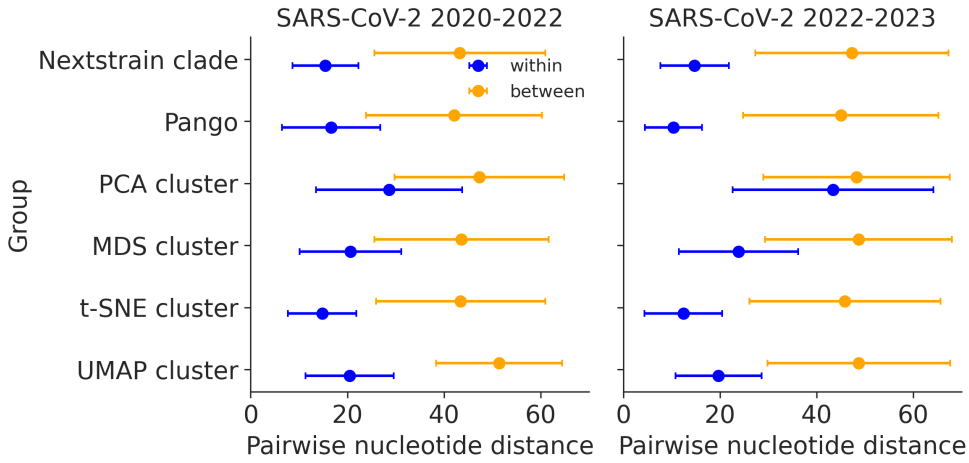
S13 Fig. UMAP embeddings for influenza H3N2 HA sequences only (top row) and HA/NA sequences combined (bottom row) showing the HA trees colored by clusters identified in each embedding (left) and the corresponding embeddings colored by cluster (right).



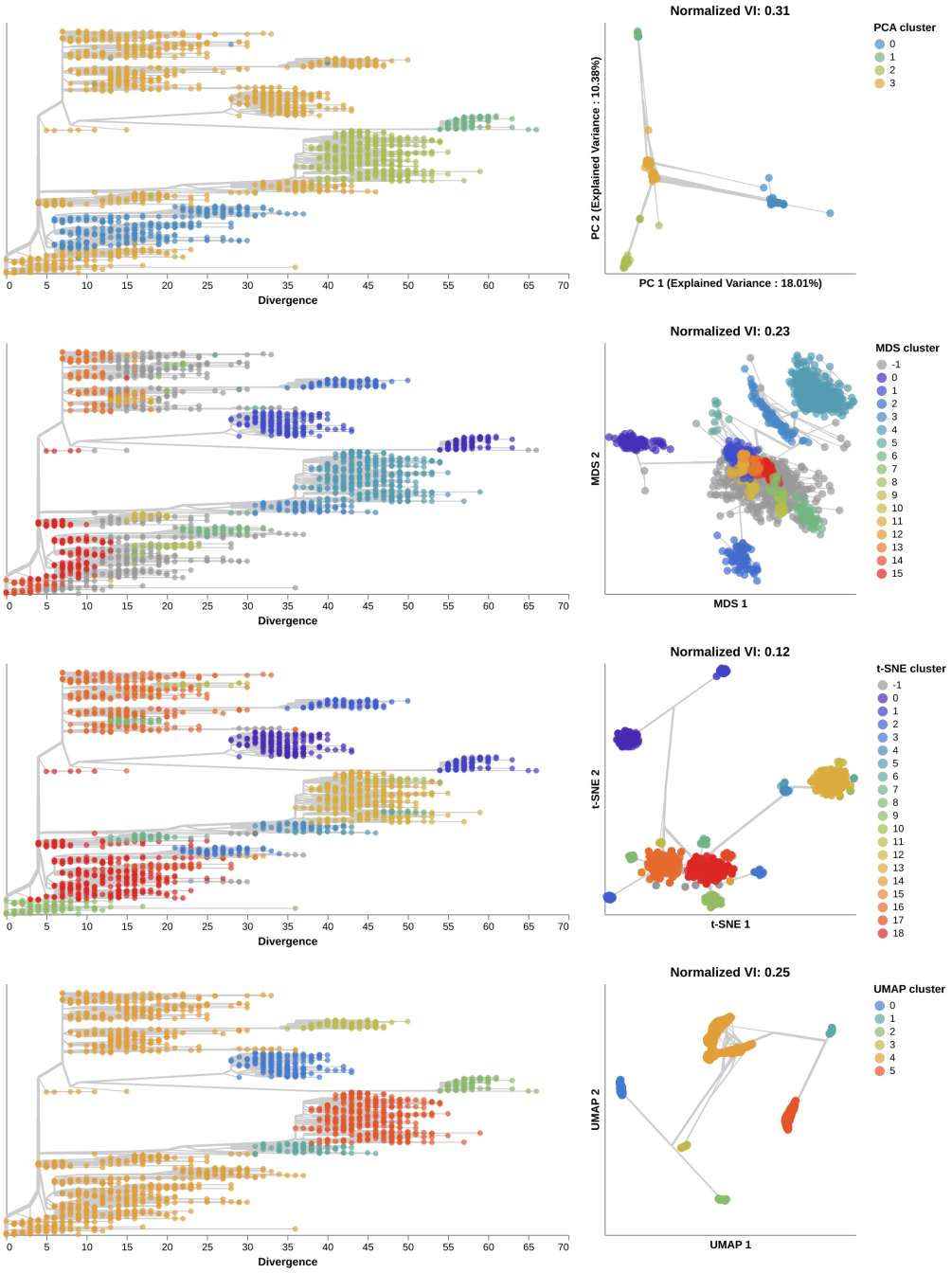
S14 Fig. MDS embeddings for early SARS-CoV-2 sequences showing all three components. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.



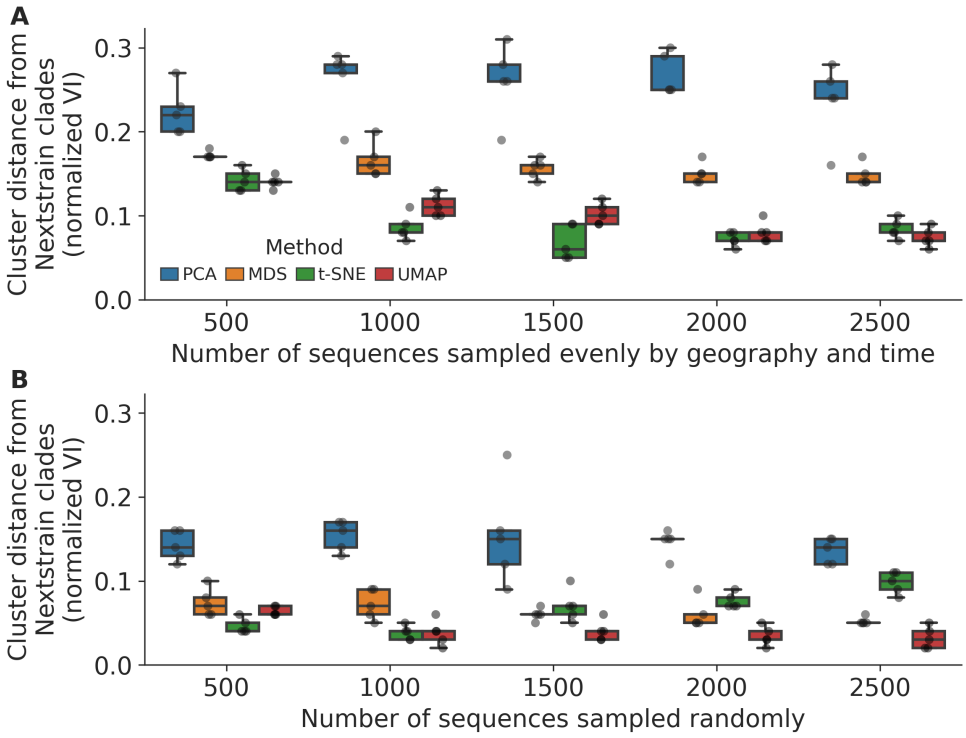
S15 Fig. Phylogeny of early (2020–2022) SARS-CoV-2 sequences plotted by number of nucleotide substitutions from the most recent common ancestor on the x-axis (top) and low-dimensional embeddings of the same sequences by PCA (middle left), MDS (middle right), t-SNE (bottom left), and UMAP (bottom right). Tips in the tree and embeddings are colored by their Pango lineage assignment. Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.



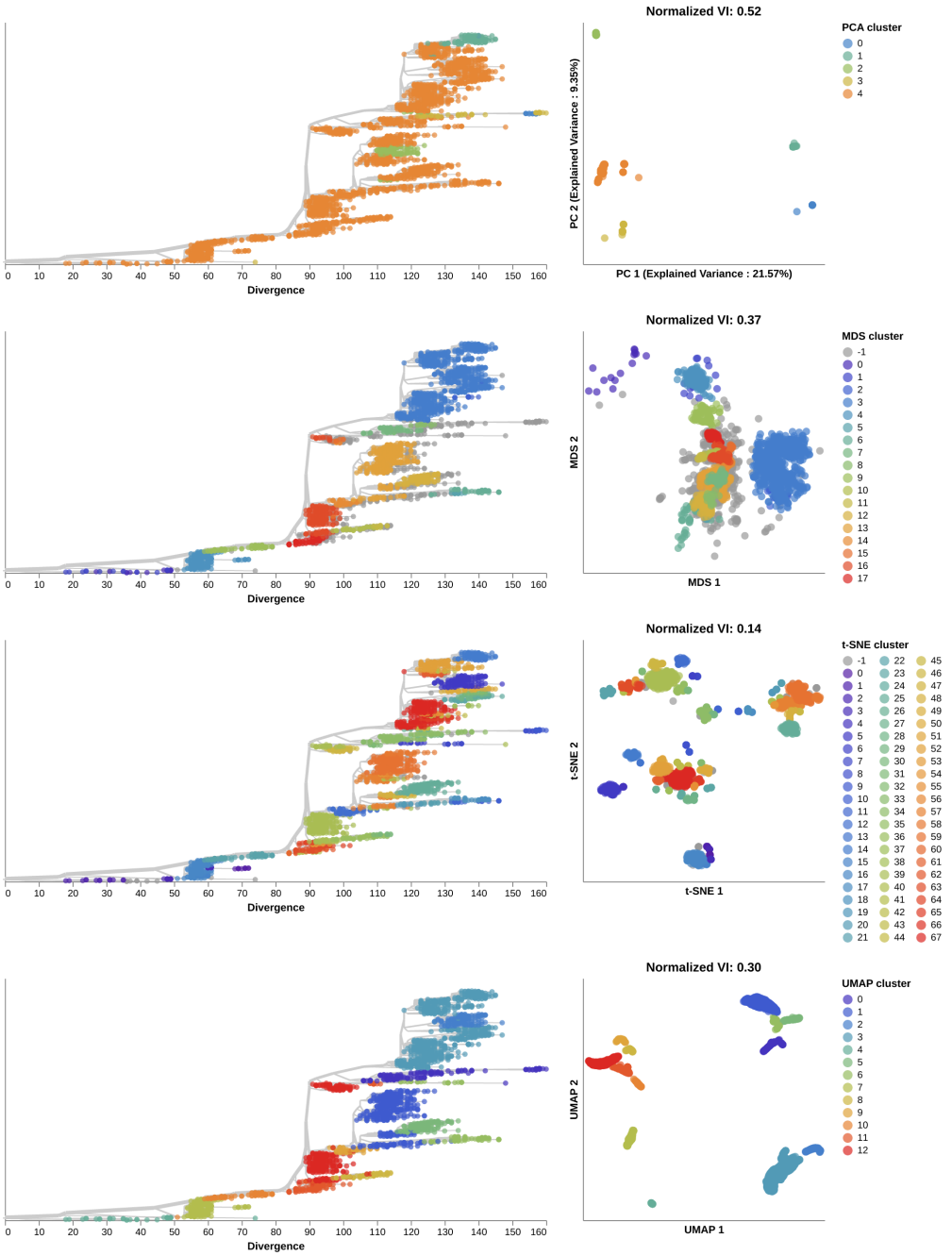
S16 Fig. Pairwise nucleotide distances for early (2020-2022) and late (2022-2023) SARS-CoV-2 sequences within and between genetic groups defined by Nextstrain clades, Pango lineages, and clusters from PCA, MDS, t-SNE, and UMAP embeddings.



S17 Fig. Phylogenetic trees (left) and embeddings (right) of early (2020–2022) SARS-CoV-2 sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Pango lineages). Line segments in each embedding reflect phylogenetic relationships with internal node positions calculated from the mean positions of their immediate descendants in each dimension (see Methods). Line thickness scales by the square root of the number of leaves descending from a given node in the phylogeny.



S18 Fig. Replication of cluster accuracy per embedding method for late (2022–2023) SARS-CoV-2 sequences across different sampling densities (total sequences sampled) and sampling schemes including A) even geographic and temporal sampling and B) random sampling. We measured cluster accuracy across five replicates per sampling density and scheme with the normalized VI distance between clusters from a given embedding and Nextstrain clades for the same samples. The even sampling scheme selected sequences evenly across region, year, and month to minimize geographic and temporal bias. The random sampling scheme uniformly sampled from the original dataset, reflecting the geographic and genetic bias in those data.



S19 Fig. Phylogenetic trees (left) and embeddings (right) of late (2022–2023) SARS-CoV-2 sequences colored by HDBSCAN cluster. Normalized VI values per embedding reflect the distance between clusters and known genetic groups (Pango lineages).